IEEE Access
Multidisciplinary Rapid Review Open Access Journal

Received February 27, 2020, accepted March 18, 2020, date of publication March 24, 2020, date of current version April 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2982932

An Autonomous Spectrum Management Scheme for Unmanned Aerial Vehicle Networks in Disaster Relief Operations

ALIREZA SHAMSOSHOARA[1], FATEMEH AFGHAH[1], ABOLFAZL RAZI[1], SAJAD MOUSAVI[1], JONATHAN ASHDOWN[2], AND KURT TURK[2]

[1]School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011, USA
[2]Air Force Research Laboratory, Rome, NY 13441, USA

Corresponding author: Alireza Shamsoshoara (alireza_shamsoshoara@nau.edu)

The work of Fatemeh Afghah, Jonathan Ashdown, and Kurt Turk was supported by the US Air Force Research Lab.

ABSTRACT This paper studies the problem of spectrum shortage in an unmanned aerial vehicle (UAV) network during critical missions such as wildfire monitoring, search and rescue, and disaster monitoring. Such applications involve a high demand for high-throughput data transmissions such as real-time video-, image-, and voice- streaming where the assigned spectrum to the UAV network may not be adequate to provide the desired Quality of Service (QoS). In these scenarios, the aerial network can borrow an additional spectrum from the available terrestrial networks in trade of a relaying service for them. We propose a spectrum sharing model in which the UAVs are grouped into two classes of *relaying UAVs* that service the spectrum owner and the sensing UAVs that perform the disaster relief mission using the obtained spectrum. The operation of the UAV network is managed by a hierarchical mechanism in which a central controller assigns the tasks of the UAVs based on their resources and determine their operation region based on the level of priority of impacted areas and then the UAVs autonomously fine-tune their position using a model-free reinforcement learning algorithm to maximize the individual throughput and prolong their lifetime. We analyze the performance and the convergence for the proposed method analytically and with extensive simulations in different scenarios.

INDEX TERMS Autonomous UAV networks, multi-agent systems, Q-learning, reinforcement learning, spectrum sharing.

I. INTRODUCTION

Several unique features of unmanned aerial vehicles (UAVs) including low cost and fast deployment, wide field of view, 3-dimensional movements, and aerial and terrestrial mapping make them very attractive for various applications such as disaster relief, military missions, wildfire monitoring, precision agriculture, and surveillance [1]–[5]. The operation of drones in such missions fully depends on the performance of the communication network among the drones and between the drones and the ground station as discussed in [6]. Unmanned Aerial Systems (UASs) demonstrate high capabilities in disaster relief operations, search-and-rescue (SAR), and aerial surveillance, however there are still several challenges toward full-scale implementation of drones in these applications [7]. One key challenge related to the application of drones in such highly-sensitive missions is to guarantee the existence of a low-latency high-throughput communication network. In certain applications namely military missions and disaster relief operations, there is a demand for high bandwidth communications to transmit sensed information to a data fusion center. Usually, the sensed information can be in the form of real-time video or high quality aerial images. The required data transmission rate depends on the dynamicity level of the operation field. For instance, the mission may involve short-term periods of time in which a very large bandwidth is required for real-time streaming that was not foreseen in the original spectrum allocation planning. In such cases, the pre-allocated spectrum to the UAV network may not be adequate to meet this demand. This need calls for new solutions to provide an additional spectrum for the fleet of UAVs during such critical missions.

In this study, we propose a solution to provide the required additional spectrum for the UAV network by considering the

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

network throughput and lifetime. In common UAVs' operation fields in rural or urban settings, there usually exists a terrestrial network which owns the licensed spectrum also known as the primary network (PN). This primary network can occasionally experience a low quality of communication due to shadowing, fading, or even the direct communication may be compromised because of the damages to infrastructure caused by natural disasters such as wildfires and earthquakes. Hence, utilizing a relay node can be beneficial for the primary or the terrestrial network to forward its messages to its legitimate receiver or improve its transmission quality by taking advantage of cooperative communication. For instance, a low-power IoT node in a remote area can benefit from a relay UAV when the direct communication to a distant destination requires a high transmission power. Our proposed spectrum sharing model exploits this opportunity to provide the additional required spectrum for the UAV networks during critical missions. In this model, one UAV serves as an aerial relay for the primary network in exchange for the required spectrum access that is borrowed for other UAVs in the network to complete their mission. Cooperative spectrum leasing models, also known as property-right models, in which the primary user as the owner of the spectrum leases a portion of its spectrum to the secondary users in exchange for some profit have been studied in several works [8]–[10]. In these models, the primary network selects the appropriate relay nodes and determines an optimal time allocation strategy for spectrum access to maximize its own benefits. However, such unilateral spectrum leasing models are not appropriate to provide the extra demanded spectrum for the UAV networks during the critical missions as the spectrum sharing strategies are solely determined by the PU.

This paper investigates a disaster relief scenario where a fleet of UAVs collect imagery data (e.g., real-time video) and transmit the data to an emergency center. We develop a hierarchical cooperative spectrum sharing model for the UAVs to secure additional spectrum access from other available network. In this model, the task of the UAVs and the most appropriate region of operation for them is assigned by a central controller. This approach divides the grid surface into multiple regions and the controller unit determines the operation region of each UAV to optimize the throughput and maximize the network lifetime considering the residual energy of each agent. The controller takes into account the level of system's dynamicity in terms of several factors such as the variations in the environment conditions due to the disaster or the primary's location to re-initiate the region assignment for UAVs if needed. Next, the UAVs perform a learning technique to find their optimal positions within their operation cell in a distributed manner. The number of UAVs is chosen based on the impacted regions. Moreover, we assumed that the task of drone is to monitor an individual region to gather video data from the impacted area and send them to the emergency center. We assumed that, we have enough drones to map them to the impacted areas and only one drone operates in a region. The UAVs are assumed to be autonomous in the sense that they can find their optimal path and location in a self-directed manner. They consider different factors such as the residual energy, position of the transmitter and the receiver and the throughput rate. A proof of convergence is provided along with simulation results to show the system performance in terms of throughput and lifetime maximization with minimizing the number of steps.

The contribution of this work is to develop a practical model for spectrum sharing in high-priority critical missions that can be locally managed by the users without the need for developing new regulations. The proposed hybrid approach utilizes both centralized and distributed techniques to find the best solution for spectrum sharing and location optimization in a reasonable amount of time. The solution guarantees the optimum throughput and lifetime for both primary and secondary (i.e., UAV) networks. We like to note that the centralized decision making is not performed constantly but as needed depending on the level of variations in the system status. Hence, the center is not considered as a bottleneck for the system performance. Moreover, this hybrid solution can enable the spectrum sharing in scenarios where the central controller (i.e., the emergency center) does not have a real-time observation of the environment, hence it can assign the UAVs to the high-priority regions and then UAVs can take into account their real-time observations of the network in order to determine their optimum position within the operation region assigned to them by the center. The center first assigns the UAVs to regions not the exact location or the cell inside the region. We assumed that center cannot fully observe the impacted regions. Therefore, the center will only determine the operation region for the UAVs based on flight distance and the residual energy of the battery. The model is scalable in the sense that adding or removing the UAVs does not impact the performance of other UAVs as they work independently. To the best of our knowledge, this work is one of the first ones to address the problem of spectrum shortage for UAV networks during critical missions.

The rest of the paper is organized as follows. Section II studies the related work regarding spectrum management in UAV networks. Section III discusses the system model, assumptions and formulation details. In Section IV, we propose two search algorithms from the emergency center view, then introduce a multi-agent Q-learning algorithm and analyze its convergence. Section V presents the simulation results and discusses the performance of the proposed methods in different settings. Finally, we conclude the study and give some future directions in Section VI.

II. RELATED WORK

While the spectrum scarcity will be a serious challenge in UAV networks given the increasing number of UAVs and the requirements of advanced wireless services, the problem of spectrum management in these networks has been barely investigated so far. mmWave communication has been discussed as an option for payload communications of UAVs as a part of 5G, but the technology is not widely available yet.

Moreover, the mmWave communication is impacted by high propagation path loss, thereby, it requires the UAVs to be equipped with high directional antennas to avoid blockage zones and maintain a LOS communication. First, we provide a brief review of some recent works focusing on the trajectory planning and deployment optimization in UAV networks based on the communication between the ground users and the UAVs. The authors in [11] considered a UAV-aided data collection scenario and developed a mechanism to optimize the trajectory, altitude, speed, and data links with the ground users to reduce the mission time of data collection from the ground users. Two approaches of segment-based trajectory optimization algorithm (STOA) and group-based trajectory optimization algorithm (GTOA) are developed to solve this optimization problem. In other works [12], [13], the authors considered a disaster relief problem, where the UAVs are used as the base stations to enhance the communication coverage for the ground users. The authors proposed two approaches to solve this problem. The first one is a centralized method to find the minimum number of required UAVs to meet the requirements for the ground users' coverage knowing the exact location of the ground users. The second solution is a distributed method that works for a given number of UAVs with no global information about the ground users. In this method, the UAVs autonomously control their motion considering the effect of virtual forces while they maintain their air-to-air communication and provide the required coverage for the ground users. In [14], the problem of spectrum allocation based on cluster formation in a cognitive IoT system is studied. The authors solved the optimization problem for the cluster size while the method aimed to maximize the formulated utility function and minimize the communication delay.

The majority of recent works for the application of UAV networks during disaster situations focus on enhancing the communication coverage for terrestrial networks using drone as temporary base stations, when the communication infrastructure is damaged. In [15], a disaster scenario is considered where the drones as the 0^{th} *responders* fly to the damaged regions before the first responders arrive at the disaster location and provide communication service to the victims. The optimal localization of the drones is determined in two phases. During the first step, a genetic algorithm (GA) method is used based on some available information from the past to initialize the location of drones. The second phase includes adapting the drones to the real-time conditions based on a local search algorithm such as hill climbing algorithm (HCA). The authors in [16] addressed an NP-hard problem of UAV network deployment to find the optimal coverage of ground users. In this paper, a multi-layout multi-subpopulation genetic algorithm is developed to find the best location of the UAVs considering the redundancy, fault tolerance, and the coverage. They compared their proposed method with traditional GA, HCA, and particle swarm optimization (PSO).

The few existing works related to spectrum sharing with drones mainly focused on the coexistence of UAV networks with cellular networks and adopted common notions of spectrum sharing such as interweave method to let the UAVs opportunistically access the spectrum holes of other communication systems, or the underlay method to allow the UAVs to utilize the spectrum of other systems while maintaining a low interference level [17]–[25]. The spectrum sensing method is not an ideal option noting the considerable energy consumption involved in searching a wide range of frequencies. More importantly, the spectrum holes are often sparse and appear on different frequencies, therefore they cannot offer continuous communication for the UAV system or require frequent changes of the operating frequency. The spectrum sharing techniques based on databases control (e.g., TV white space) only allow a low level of transmission power for unlicensed users, and allocate a wide and static protection zone around the incumbents. One common drawback of these conventional spectrum sharing methods is that the spectrum owners are oblivious to the presence of the devices seeking for spectrum, but a dynamic and efficient practical spectrum sharing model cannot be implemented unless different users including the spectrum owners and the ones looking for spectrum interactively cooperate with one another.

A primary model for cooperative spectrum sharing in UAV networks where the UAVs are divided into two clusters of sensing and relaying UAVs is proposed in [26]. In this model, the UAVs are assumed to be located in fixed positions with no movement and a multi-agent reinforcement learning-based solution was developed to find the best task allocation for each UAV in a distributed manner. While this solution can be applicable to scenarios where the communication among the UAVs is not available or reliable, such fully distributed approach for the task allocation with no message exchange among the agents does not guarantee an optimum outcome. In [27], the problem of cooperative spectrum sharing between a UAV network and a terrestrial network was considered, where a fully centralized approach is proposed serving for both task allocation between the sensing and relaying UAVs, and trajectory optimization. The authors assumed that the emergency center has full control over the UAVs in terms of the task assignment and movement. This problem was solved using a "team Q-learning" algorithm for a multi-UAV scenario. Although the proposed method is able to find the best solution considering the UAVs' locations and assigned tasks; however, it takes a considerable amount of time for the learner to get experience in large grids. Therefore, this centralized model is not scalable and it also relies on a full real-time knowledge of the central controller about the network and the UAVs' status.

In this paper, we propose a hierarchical joint task allocation and path planning method that offers an scalable and reliable solution for cooperative spectrum sharing for UAV networks. In this model, first the emergency center determines the operation region and the proper task of each

UAV using a search algorithm or a bipartite graph matching method [28], [29]. Then, the UAVs determine their optimum location within their operation region given their real-time observation of their environment. The UAV coordination for all agents is performed using Reinforcement Learning (RL) to achieve a common goal. In fact, by using the hybrid approach, the nature of the problem which was a Multi Agent RL is simplified to several single agent RL sub-problems.

III. SYSTEM MODEL

Let us assume a fleet of N autonomous UAVs with the mission of transmitting high data rate surveillance data such as video or images to a pre-defined emergency center. We assume that all UAVs are autonomous in a way that they can control their path in their region and determine their optimal location autonomously. The emergency center can be considered in a form of a ground station or a High Altitude Platform (HAP) in a fixed location. This emergency center has a prior knowledge about the impacted areas and their level of priority, but it may not necessarily have a real-time observation of the operation field.

It is assumed that there exists a licensed pair of terrestrial transmitter-receiver willing to share their spectrum with the UAV network as they suffer from a low quality of communication or they are interested to improve the communication quality through relaying service by the UAVs. We assume that the information related to the availability and location of such primary user is made available to the central controller. This information can be either released by the cellular network or the UAVs can gather the required information about the location of the primary users which experience a low quality of service from their network. In [30], a Bayesian approach is developed to localize the ground users using their received signal strength indicator (RSSI). A similar approach can be utilized in our model to locate the primary users which are in need to receive a better quality of service through spectrum leasing with the relaying UAVs. For the sake of simplicity, we assume that the primary's transmitter and receiver are too far from each other that the direct communication is not feasible or efficient for them [31]–[33]. Hence, they need a relay to forward their messages for them. We should note that the proposed spectrum sharing method can be also utilized in scenarios where the direct link between the primary's transmitter and receiver exists but the PU takes part in leasing its spectrum in exchange for diversity gain using cooperative relaying methods. The pair of PU's transmitter and receiver are located in a fixed location on the ground and the UAVs (the secondary users) are assumed to hover in a fixed altitude during their mission. The proposed model can provide a spectrum sharing solution between the drones and any terrestrial users which operate on a frequency range that is supported by the drones. However, noting the wide-area coverage, and security of mobile networks, the cellular networks are proven to be the most reliable choice to serve the UAV networks, in particular during critical missions [34]. Therefore, in this paper we consider the ground users as LTE users. We like to note that the proposed model is different from the current trend of coexistence of UAV networks with the LTE network, where the drones are served by the side lobes of LTE base station antennas. During the disaster situations, it is very likely that the cellular base stations may be damaged. Moreover, the interference caused by the drone communications to the cellular networks needs to be further investigated. In our model, we propose a solution that the LTE users intentionally lease a portion of their time access to their spectrum to the UAV networks, in exchange for cooperative relaying from these flying relays when the local LTE base stations are damaged.

It is assumed that the UAVs are provided with a dedicated bandwidth for Control and Non-Payload Communication (CNPC) to exchange signaling and controlling information and also with a limited bandwidth for payload communications which may not be adequate to deliver real-time transmission of high-resolution data such as video and image during disaster relief operations. Therefore, the UAV network may require additional temporary spectrum access. To meet this demand, one UAV in the network is selected by the emergency center to act as an aerial relay for the terrestrial network, while the rest of the UAVs can take advantage of the spectrum access provided by the primary network to transmit their collected information to the emergency center. The PU grants half of its time of spectrum access to the UAV network and in exchange one selected UAV assists the PU to deliver its packet to the legitimate receiver. In summary, the UAVs can be categorized into two sets of *sensing* and *relaying* UAVs. One UAV is selected as a relay and $N-1$ operate as the sensing UAVs.

We consider a hybrid scheme consisting of both the centralized and autonomous control scenario in which the task of the UAVs is determined by the emergency center while they have the capability to decide for their actions in terms of the mobility to maximize their throughput and lifetime. Thus, the controller does not require continuous updates on the environment status to determine the real-time location of the UAVs, rather it assigns the sensing UAVs to high-priority regions based on its prior knowledge of the environment in the beginning of the mission. Then it revises this allocation as needed if the level of dynamicity such as PU's location or the fire's growth exceeds some pre-defined threshold.

Figure 1 demonstrates a sample scenario with three sensing and one relay UAVs. The UAVs are assumed to be located in a plane (not the emergency center). The emergency center identifies the high priority impacted areas and clusters these regions to multiple non-overlapping operation fields. The size of operation regions depends on several factors including the application type, the number and type of the UAVs, the shape and dimensions of the impacted areas. The emergency center assigns the UAVs to the optimal operation field for them based on the residual energy and number of hops they have to fly to reach the intended region and then the UAVs can fly in their operation region and find their optimum location within their region by taking one of the actions of {Up , Down , Left , Right , Stay}.

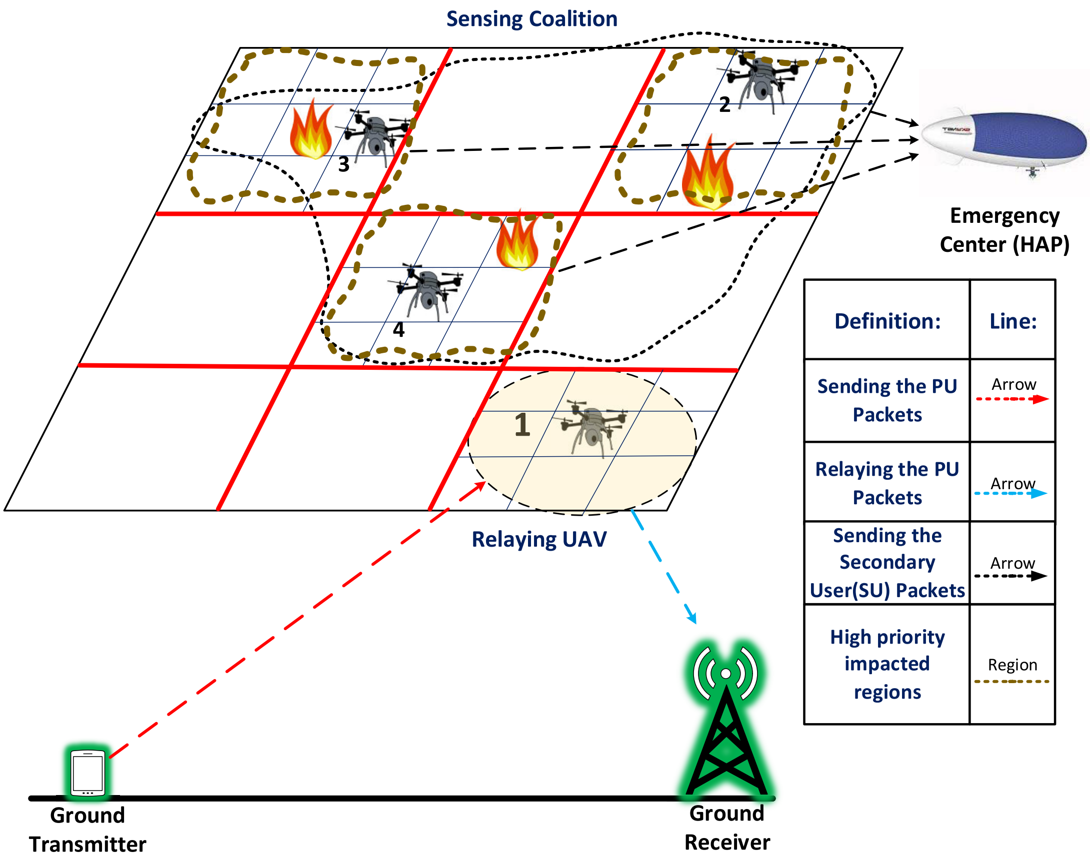


FIGURE 1. A sample scheme of the proposed spectrum sharing between a UAV network and a terrestrial licensed network. Those red lines demonstrate the clustering of the whole area after prioritizing the impacted area and also the size of each region. The regions in the red lines are non-overlapped.

In each time slot, the channel state information (CSI) between all the nodes in the network including the emergency center, the ground users, and the UAVs is determined based on the distance between the chosen source and destination followed by a simple LoS model. The CSI parameters between the PU and the relaying UAV follow the slow Rayleigh fading and are available at the transmitter based on common channel estimation techniques [35]–[37].

In this work, we defined the CSI parameters as follow: h_{PT,U_i} expresses the CSI between the primary transmitter and i^{th} UAV, $h_{U_i,PR}$ describes the channel between the i^{th} UAV and the primary receiver, $h_{U_i,E}$ carry the information for channel between i^{th} UAV and the emergency center. The noise of all channels are modeled at the receiver side with normally distributed symmetric complex values, $Z \sim CN(0, \sigma^2)$. We assume that all nodes including the UAVs and the ground users have a constant arbitrary value for the transmission power. These arbitrary values are different for the UAVs, the source, and the primary transmitter. While several works such as [38]–[40] focused on optimizing the power consumption, we consider a constant transmission power; however, we address the energy consumption rate by minimizing the number of movements in the plane and each region.

We consider half-duplex transmission for the primary network. In the first half of the time slot, the primary transmitter sends its packet to the relaying UAV and then the UAV forwards the packet to the primary receiver. The sensing UAVs transmit the gathered information such as video or sensed data to the emergency center during the second half of the time slot. Figure 2 illustrates the time allocation between two networks in each slot.

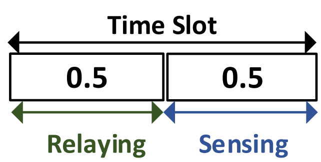


FIGURE 2. The allocation of each time slot between the relaying and sensing tasks.

Different relaying schemes such as amplify and forward (AF) and decode and forward (DF) can be utilized at the relay UAV [41]–[44]. Here, we utilize the AF scheme, hence the throughput rate for the primary user can be achieved as [45]:

$$R_{PU} = \frac{1}{2}\log_2\Big(1 + P_{PT}|h_{PT,PR}|^2 + \frac{P_{PT}|h_{PT,U_i}|^2\, P_{U_i}|h_{U_i,PR}|^2}{\sigma^2 + P_{PT}|h_{PT,U_i}|^2 + P_{U_i}|h_{U_i,PR}|^2}\Big), \quad (1)$$

where the transmission power for the primary transmitter and UAV i are denoted by P_{PT} and P_{U_i}. Background noise power at the receiver is σ^2, and let i denote the index for the i^{th} (UAV). Due to the long distance between the PT and PR, $h_{PT,PR}$ is zero, hence we can eliminate this term from the throughput rate ($|h_{PT,PR}| \simeq 0$).

The throughput for each sensing UAV can be calculated as [46], [47]:

$$R_{SE_i} = \frac{\lambda_i}{2} \log_2(1 + \frac{P_{U_i}|h_{U_i,E}|^2}{\sigma^2}), \tag{2}$$

where λ_i is the portion of time that is allocated to i^{th} UAV transmission considering the priority of its operation field.

The achievable rate at the emergency center, R_{SE}(Multi-UAV) can be obtained as

$$\begin{aligned} R_{SE}(\text{Multi-UAV}) &= \sum_{i \in U_E} R_{SE_i} \\ &= \sum_{i \in U_E} \frac{\lambda_i}{2} \log_2(1 + \frac{P_{U_i}|h_{U_i,E}|^2}{\sigma^2}), \end{aligned} \tag{3}$$

where $U_E = \{E_1, E_2, \ldots E_{N-1}\}$ denotes the sensing set.

In Figure 3, a sample case with $(N-1)$ sensing UAVs and one relaying UAV is shown. Figure 3 shows channel state information between different nodes in this scenario.

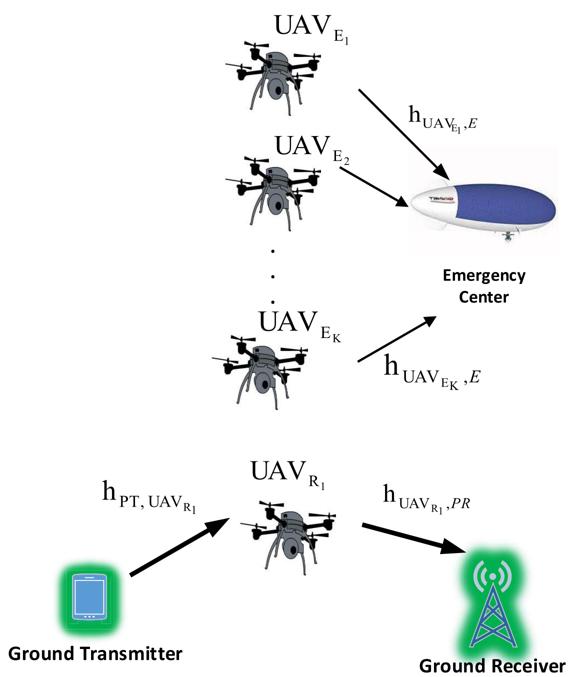


FIGURE 3. Channel state information for the relaying UAV and sensing coalition.

IV. PROPOSED ALGORITHM

The proposed algorithm for spectrum sharing includes two steps: (i) the search algorithm which picks the best region for the primary network based on the throughput metric and assigns UAVs to the prioritized regions and the primary region to prolong the network lifetime, and (ii) the reinforcement learning algorithm performed by the individual UAVs to gain experience on the states and actions to decide movement actions wisely and maximize the individual throughput and lifetime. In the simulation results, we observed both individual throughput and the network throughput which is the summation of all UAVs' throughput.

A. TASK ALLOCATION AND REGION ASSIGNMENT BY THE EMERGENCY CENTER

We assume that at the beginning of each round of optimization, the UAVs are located at random cells across the region with different random initial energy. The emergency center determines the number of regions in the grid-operation field knowing the number and type of available UAVs. It also takes into account its prior knowledge of the impacted field to identify the high priority regions. After determining the high-priority regions and their IDs, the emergency center identifies a proper primary network within the operation field and assigns the best relay UAV to the region close to the PU in order to maximize the throughput rate of the PU. Selecting the region for the relaying UAV depends on the Euclidean distance between the region and the primary receiver using the search algorithm described in (4).

$$\text{UAV}_P = \underset{u \in \mathcal{U}}{\text{argmax}} f(u) := \{u | u \in \mathcal{U} : E(u) - d(u)\Psi\}, \tag{4}$$

where, UAV_P is the index of the chosen UAV to serve the primary network, $\mathcal{U}$ is the set of all UAVs, $E(u)$ is the initial energy for UAV u, $d(u)$ is the number of steps or hops that UAV u should travel to reach the center of the primary region, and Ψ is the energy consumption rate per movement. Finding the u guarantees the longest lifetime for the primary network. After assigning the primary UAV to the primary region, the UAV's initial energy will be decreased due to the flight distance.

Next, the emergency center chooses arbitrary number of regions based on the disaster operation and the priorities for the sensing area. To utilize all UAVs, we assume that the arbitrary number of regions is equal to number of sensing UAVs. Afterward, it assigns the remaining UAVs to the high-priority regions such that the residual energy after the flight distance from the initial location to the region is maximized. The emergency center makes a preference list including the sorted indices of the regions for each UAV. Next, it allocates the UAVs to the regions based on the preference value. If two or more UAVs have the same preference value for one specific region, then the emergency center assigns the UAV with higher remaining energy to this region and allocates the second UAV to its next preference region. Also, in case that two or more UAVs have exactly the same preference values and same amount of residual energy, the center assigns the UAVs to regions by random because there is no difference between these UAVs. Expression (5) shows the preference array for the sensing UAV i:

$$\text{R}_{S_i} = \underset{r \in \mathcal{R}}{\text{argsort}}\, g(r) := \{r | r \in \mathcal{R} : E_i - d_i(r)\Psi\}, \tag{5}$$

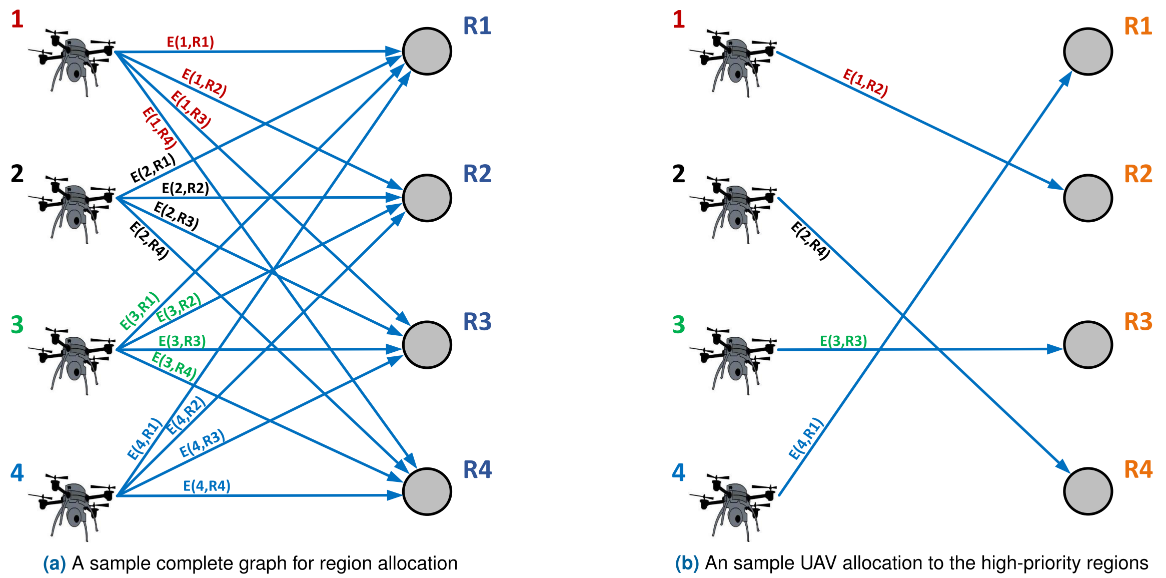


FIGURE 4. An example for matching the UAVs to the high priority regions.

where, $\operatorname{argsort}_{r\in\mathcal{R}} g(r)$ is the function to order a list in a descending order and returns the index of those values, R_{S_i} is the preferred regions for UAV i, $\mathcal{R}$ is the set of high-priority regions for the secondary network. E_i is the initial energy for UAV i at its first location. $d_i(r)$ is the distance of UAV i from its initial location to the region r.

Figure 4a shows a complete graph with the sensing UAVs and the region centers as the vertices and the remaining energy after the flight distance as the edges. Figure 4b shows an example of the allocated UAVs to the regions based on the preference regions in (5). This allocation mechanism guarantees the longest lifetime for each UAV. Based on this ranking criteria, the goal of region selection is to have an optimal throughput rate and the purpose of the UAV allocation is to get the longest lifetime for both networks considering the energy. After allocating the UAVs to the regions, each UAV utilizes a Q-learning algorithm to find the best cell in each region.

B. LOCATION OPTIMIZATION AT INDIVIDUAL UAVs USING REINFORCEMENT LEARNING

In applications where training data-sets are available to learn the optimal behavior, the artificial neural network-based (ANN) approaches are commonly utilize to learn the optimal solution. However, in applications such as our problem that lack the required training data-sets or knowledge about the optimal solution, reinforcement learning-based methods can offer a trial-and-error approach to determine the solution. ANN-based methods are limited to the given training data-sets for the optimal solution and the obtained solution cannot be better than the provided samples. However, a key advantage of the RL-based approaches is that the agent can learn to behave better than an expert in the problem.

In this study, we break down the Multi-Agent Reinforcement Learning (MARL) into several single RL sub-problems. Each UAV is considered as a single agent which is independent of other UAVs. Each UAV is connected to its region with action and perception [48].

Based on the aforementioned system model in Section III, each UAV operates in an individual region area with no interaction with other UAVs. The states, actions, and rewards are exclusive for each UAV. Taking an action for a specific UAV does not impact the state and reward for other UAVs. As a result, a simple Q-learning is used to find the best location for each UAV. Each UAV monitors its behavior and decides based on its experience.

The decision making for each UAV is modeled by a Finite Markov Decision Process (FMDP) framework to construct the decision making in the discrete time stochastic control process. Each UAV only observes its own state refers to its location within its allocated region. Assume that $S_t^i \in \mathcal{S}^i$ defines the state for the i^{th} UAV at step t, $\mathcal{S}^i$ is the set of all states for UAV i. Taking an action such as $a_t^i \in \mathcal{A}^i$ for i^{th} UAV changes its state from S_t^i to S_{t+1}^i. Choosing an action for each UAV defines its behavior which follows the policy $\pi(S)$ based on (6).

$$\pi(S_t) = \operatorname*{argmax}_{a_t \in \mathcal{A}(\mathcal{S}_t)} \pi(S_t, a_t), \tag{6}$$

The default action set $\mathcal{A}$ includes four actions for movement and one action for staying at the previous location $\mathcal{A} = \{\uparrow (0), \downarrow (1), \leftarrow (2), \rightarrow (3), -(4)\}$. Depending on the UAV's location, some particular actions may be prohibited. For instance, if the UAV is located on the region's edge side, taking an action which results to leave the region is prohibited. Taking a new action and altering into a new state (S_{t+1}) updates the reward value $r_t \in \mathcal{R}$ in respect to (9). Following the optimal policy based on (6) guarantees the action set which results into expected reward maximization $\mathbb{E}(\sum_{t=0}^{\infty} \gamma^t r_t)$, where $\gamma \in [0, 1)$ is the discount factor. Individual UAV forms distinct MDP including a 4-tuple which can be shown as $(\mathcal{S}, \mathcal{A}, P_a, r_a)$, where $\mathcal{S}$ is finite set of states for each UAV. Since all the regions sizes are the same and the UAVs are independent, all UAVs consider the same state set, $\mathcal{A}$, a finite action set which depends on the UAV's location in the region. $P_a : S_t \times A \times S_{t+1} \rightarrow [0, 1]$ is the state probability function which stands for the randomness in transition between the S_t and S_{t+1}, $r_a : S \times A \times S_{t+1} \rightarrow \mathbb{R}$ is the gained scalar reward for an individual UAV if it takes an action a and changes its state from S_t to S_{t+1}. This reward introduces the effect of the immediate action. However, it does not interpret the long-term effect of the action. Hence, the UAV needs to maximize the long-term reward. One approach is using the optimal action-value function and computing that with the Q-learning algorithm [49], [50] which satisfies the Bellman optimality equation. All UAVs update their Q-values based on:

$$Q(S_t, a_t) = (1-\alpha) \times Q(S_t, a_t) + \alpha[r_a(S_t, S_{t+1}) + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a')], \quad (7)$$

where, $\alpha \in (0, 1]$ is the learning rate for the agent. $Q(S_t, a_t)$ is the action-value for the current state which depends on the α, current reward, γ, and the $\max_{a' \in \mathcal{A}} Q(S_{t+1}, a')$. Last term stands for the maximum action-value among all values for the next state S_{t+1}. While (7) defines the Q table and values for one UAV, (8) introduces the updating matrix for all agents:

$$\begin{bmatrix} Q_1(S_t^{(1)}, a_t^{(1)}) \\ Q_2(S_t^{(2)}, a_t^{(2)}) \\ \vdots \\ Q_N(S_t^{(N)}, a_t^{(N)}) \end{bmatrix} = \begin{bmatrix} 1-\alpha^{(1)} \\ 1-\alpha^{(2)} \\ \vdots \\ 1-\alpha^{(N)} \end{bmatrix} \cdot \begin{bmatrix} Q_1(S_t^{(1)}, a_t^{(1)}) \\ Q_2(S_t^{(2)}, a_t^{(2)}) \\ \vdots \\ Q_N(S_t^{(N)}, a_t^{(N)}) \end{bmatrix} + \begin{bmatrix} \alpha^{(1)} \\ \alpha^{(2)} \\ \vdots \\ \alpha^{(N)} \end{bmatrix} \cdot \begin{bmatrix} r_a^{(1)}(S_t^{(1)}, S_{t+1}^{(1)}) + \gamma \max_{a'^{(1)} \in \mathcal{A}^{(1)}} Q(S_{t+1}^{(1)}, a'^{(1)}) \\ r_a^{(2)}(S_t^{(2)}, S_{t+1}^{(2)}) + \gamma \max_{a'^{(2)} \in \mathcal{A}^{(2)}} Q(S_{t+1}^{(2)}, a'^{(2)}) \\ \vdots \\ r_a^{(N)}(S_t^{(N)}, S_{t+1}^{(N)}) + \gamma \max_{a'^{(N)} \in \mathcal{A}^{(N)}} Q(S_{t+1}^{(N)}, a'^{(N)}) \end{bmatrix}, \quad (8)$$

where N is the number of agents (UAVs), also we assumed that all UAVs have the same learning rate and discount factor. The Q-tables are distinct for each UAV and the UAVs take actions simultaneously.

To define the reward function for the agents based on their actions in the previous time slots, we assume that both the primary receiver and emergency center report their received throughput to the UAVs. Since the goal of our model is to maximize the throughput, network lifetime, and minimize the energy consumption, the reward function consists of both the gained throughput and remaining battery of the node. The reward function compares the last two consecutive throughput rates and remaining energy to give an award to UAVs. Since the UAVs are located in different regions, their reward functions are independent of each other and there is no correlation between them. (9) defines the reward function for each UAV:

$$Reward(i) = \begin{cases} \beta_1 & \text{if } (R(t) > R(t-1)) \text{ and } (E(t-1) - E(t) = \Psi) \\ \beta_2 & \text{if } (R(t) = R(t-1)) \text{ and } (E(t-1) - E(t) = \psi) \\ \beta_3 & \text{if } (R(t) < R(t-1)) \text{ and } (E(t-1) - E(t) = \Psi), \end{cases} \quad (9)$$

where, $R(t)$ is the throughput rate at timeslot t, β_1, β_2, and β_3 are rewards based on the utility and remaining energy values. i denotes the index for i^{th} UAV. Ψ is the energy consumption rate when a UAV changes its location and performs data transmission. ψ is the energy consumption rate when a UAV stays at its location and performs data transmission. We also assume that the energy consumption due to the UAV's mobility is more than the energy used for data transmission, as it is usually the case in UAV networks. The reward is obtained based on the throughput rate and since the throughput depends on the distance between any source and destination, the UAVs' goals are to find a proper location for themselves in each region to maximize their throughput and minimize the energy consumption.

In each step, the UAVs are awarded with different reward values based on (9) which follows three possible options: first, the UAV changes its position and it improves the throughput rate too. But it costs energy to change the cell in the region. In this case, the UAV earns β_1 as a positive reward. Second, the UAV stays at its position and keeps its previous throughput rate and saves its energy, in this case, the UAV is granted with β_2 as the reward. Third, the UAV changes its location and receives less throughput rate and it also looses its energy because of the movement. In this case, the UAV is punished with β_3 as a negative value.

For the action selection process, we implement the ϵ-greedy exploration for all UAVs with the constant ϵ. Hence, each UAV chooses a random action with the probability of ϵ and it chooses the best action with the probability of $1-\epsilon$ to find the best action based on updated values in the Q-table. We like to note that in team Q-learning, a single learner decides for all agents in a random or a greedy manner. However, in multi-agents such as Nash Q-learning or the method in this paper, the UAVs act independently. For instance, some UAVs behave randomly while others choose greedy actions [51]. The utilized action selection method is shown in 10:

$$a_t^{(i)} : \begin{cases} \text{random Actions}, & \text{rand(i)} < \epsilon, \\ \operatorname{argmax}_{a^{(i)}} Q_i(S_t^{(i)}, \mathcal{A}), & \text{o/w}, \end{cases} \quad (10)$$

where $i \in \{1, 2, \ldots, N\}$ is the index for each UAV and $\mathcal{A}$ is the available actions set for the i^{th} UAV when it is located in state $S_t^{(i)}$.

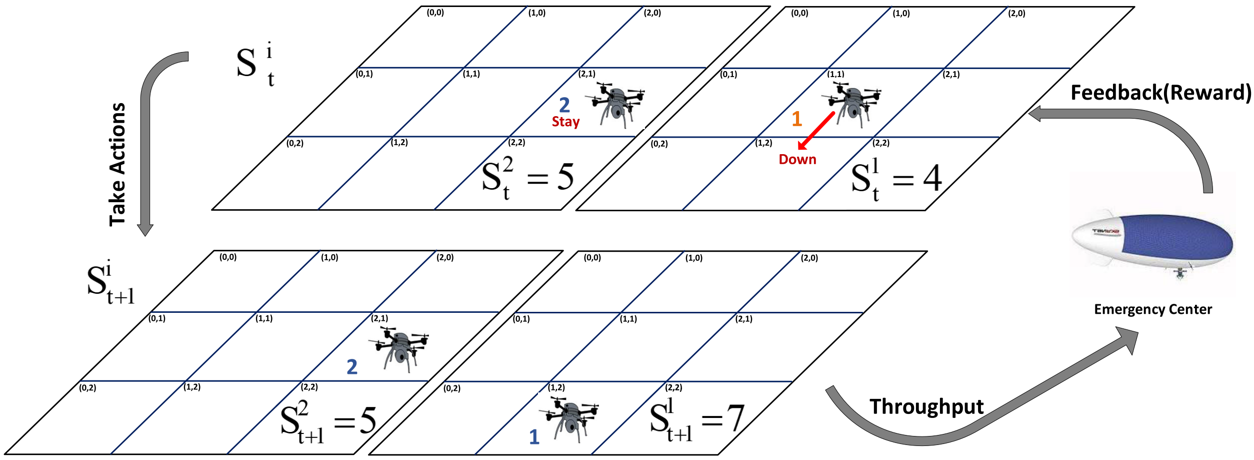


FIGURE 5. An example of RL-based decision making process for 2 UAVs in a 3 × 3 regiona.

Figure 5 demonstrates an example of the learning procedure for two UAVs based on the feedback reward, the taken actions, and the states. UAV 1 changes its location based on the taken action. UAV 2 stays at its location. These processes bring a new state for UAV 1, based on the *Down* action, it flies from state 4 to state 7. Also, UAV 2 keeps its previous state at location 5. Based on the received throughput for both the primary receiver and the emergency center as feedback, the rewards are assigned to the UAVs. Algorithm 1 summarizes the search algorithm to find the best primary region and the UAV allocation by the emergency center, also it abstracts the requiring steps for the Q-learning in the RL process for all UAVs.

C. CONVERGENCE ANALYSIS OF THE RL ALGORITHM

Since the UAVs are independent, they have different Q-tables. Let $Q_i^*(S, a_i)$ denote the optimal Q-value for the i^{th} UAV in algorithm 1.

Theorem 1: The utility functions for the system which are defined by formulas (1), (3), and (9) are bounded and finite.

Proof: Considering (3), terms $|h_{S,U_j}|^2$ and $|h_{U_j,E}|^2$ depend on the location of UAVs. Since, the emergency center and the primary users are placed in a fix location and the gird size is bounded for definite positions, then R_{SE}(Multi-UAV) is bounded for definite numbers. The same approach can be used for R_{PU}. As a result, both (1) and (3) are bounded. Also, the reward function is defined based on β_1, β_2, and β_3 which are finite values.

Theorem 2: If the learning rate is bounded between 0 and 1 ($0 < \alpha \leq 1$) for all $(S, A) \in \mathcal{S} \times \mathcal{A}$ which requires all states-actions to be observed for infinite times, then:

$$\sum_{i=1}^{\infty} \alpha = \infty, \quad \sum_{i=1}^{\infty} \alpha^2 < \infty \tag{11}$$

Proof: To prove Theorem 2, auxiliary results from stochastic approximation are needed as provided in [52], [53].

Lemma 1: The Q-learning algorithm in algorithm 1 with the update rule from (8) is converging to the optimal $Q_i^*(S, a_i)$ with probability one (w.p.1) if the utility function for the system is bounded, the state and actions sets are finite, and $\sum_{i=1}^{\infty} \alpha = \infty$, $\sum_{i=1}^{\infty} \alpha^2 < \infty$ for $0 < \alpha \leq 1$ [52], [54]:

$$\lim_{t \to \infty} Q_{t,i}(S, A) = Q_i^*(S, a_i) \tag{12}$$

where, t is the time step in each episode.

Table 1 compares the number of states and the Q-table size for both methods of team learning [27] and our proposed method in this paper for a sample size of 6×6 with 2 agents. Since, the HAP center assigns the task for UAVs, the algorithm has one less action compared to [27]. Also, we assumed that the whole grid size of 6×6 is divided into 4 regions size of 3×3, and each UAV operates in one region. The team Q-learning approach needs more space size for the Q-table since the size increases exponentially regarding the number of drones; however, the proposed method requires less space in each Q-table since there is only one UAV in each region as a result to total number of states is equal to the summation of all states.

TABLE 1. Comparison of the action and state space and the size of Q-table between the proposed method and the team Q-learning approach developed in [27] for a system model of a 6 × 6 grid.

	Sample size	
Algorithm	Proposed Method	Team Q-learning [27]
Actions	5	6
# of Agents	2	2
State	$2 \times (3 \times 3) = 18$	$(36)^2 = 1,296$
Q-Table Size	$2 \times [(3 \times 3) \times 5] = 90$	$(36 \times 6)^2 = 46,656$
Reward(Description)	Difference of throughput and energy	Difference of throughput
Reward	$R(S, A, S') \to \mathbb{R}$	$R(S, A, S') \to \mathbb{R}$

Algorithm 1 Region Assignment, UAV Allocation, and Q-Learning for RL Framework

```
initialization:
Set initial parameters and conditions
Select the primary region by the emergency center
Allocate the operation regions for sensing UAVs by the
emergency center noting the region priorities
Decrease the energy after UAV allocation
Set all initial parameters for the RL
for all Runs do
    Initial all Q-Tables, utility functions, actions, tasks,
    and all arrays to zero
    for all Episodes do
        Set the location for the UAVs to the initial
        Set the energy for the batteries to the initial
        energy
        if Variations in the environment (dynamicity
        level) > τ then
            re-initiate the controller process for the
            UAVs and regions
        end
        for all t < Step Size in the Grid do
            for all the UAVs do
                Get the State from the updated location
                if Random < ε then
                    Choose a random action
                    Update the location
                else
                    Choose an action based on best
                    Q-value from the Q-table
                    Update the location
                end
                Update the energy after flight
                Calculate the utilities and throughput
                Update the energy after transmission
                Calculate the reward
                Update the Q-Table
            end
        end
    end
end
```

V. SIMULATION RESULTS

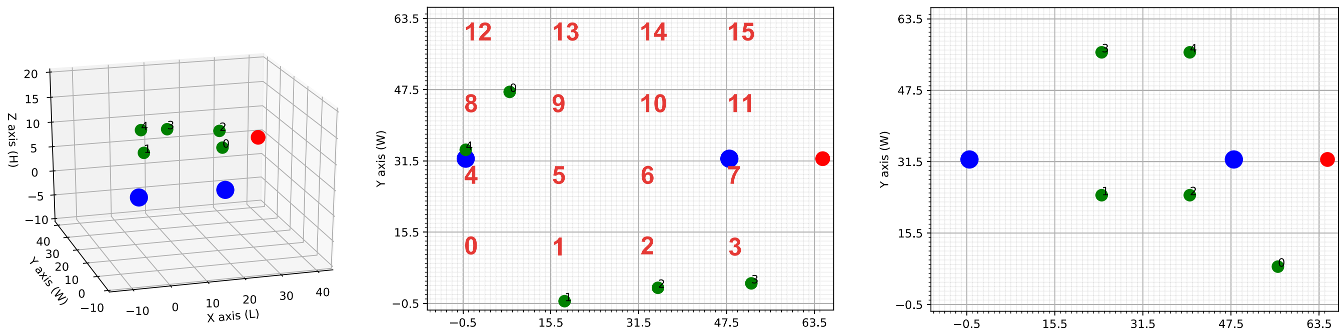
In this section, we consider different scenarios for different gird sizes and system conditions to evaluate the performance of our proposed method. First, a pair of ground-based transmitter-receiver as the primary users and an aerial-based emergency center (e.g. an HAP) are located in random locations. The whole area of coverage is surfaced into $L_1 \times L_2$ tiles. The drones are assumed to be located in arbitrary locations at the beginning of the optimization round. They can move in four directions of {Up , Down , Left , Right} or stay in the same location. The whole grid size is divided into a pre-defined number of regions. $h_{i,j}$ defined the CSI parameters between nodes i and j is calculated based on a LoS model with a known propagation loss factor. We assume that the loss factor in our scenario is -2. LoS model considers the 3 dimensional Euclidean distance between the nodes. β_1, β_2, and β_3 (the rewards in (9)) are chosen arbitrary based on the experiment in the simulations. Random values between 4000 and 5000 Jules are chosen as the initial battery value for the UAVs. The transmission powers are chosen based on $P_{PT} = 10\,mW$ and $P_{U_i} = 20\,mW$. The mobility consumption rate (Ψ) and the transmission consumption rate (ψ) in each time slot are $10.0\frac{\text{J}}{\text{Mobility}}$ and $0.5\frac{\text{J}}{\text{transmission}}$, respectively. $\sigma^2 = 1nW$. RL parameters α, γ, and ϵ are chosen as 0.1, 0.3, and 0.1, respectively unless mentioned explicitly otherwise. For the RL algorithm, we assume that the rate of the exploration-exploitation is fixed meaning that the UAVs with a constant probability (ϵ) choose the actions based on the best action-state Q-values otherwise they do it randomly.

Since each UAV is located in a separate region, the state-space is $(R_1 \times R_2)$, which R_1 and R_2 are the dimensions for the region area. All regions have the same dimension during the simulation unless the emergency center changes them at the beginning of each run. The number of UAVs is 5 in all iterations and episodes. The emergency center applies a simple search algorithm to find the best primary region and the relay UAV. Also, it uses the bipartite graph matching to allocate UAVs to the regions. The time complexity to find the best primary region is $O(M)$, where M is the number of regions and the time complexity for finding the best relay UAV is $O(N)$, where N is the number of UAVs. Next, the emergency center assigns the sensing UAVs to the high priority-regions with the time complexity of $O(N)$ and utilizes the bipartite graph matching. It matches the UAVs to the high-priority regions with the time complexity of $O((N-1)^2) \sim O(N^2)$.

20 external iterations are considered to improve the accuracy of the simulation and avoid intolerance of the initial values and a biased behavior. In each iteration, the UAVs run 40 episodes to fill the gaps of Q-values in their Q-tables. The Q-tables are initialized with zeros and the final Q-values in each Q-tables is the initialization for the next episode. The number of steps in each episode is assigned based on the experiment and the number of states in each region. Table 2 brings the parameters and timing values for all the executed simulations. Timing values are calculated using a system with AMD RYZEN 9 3900X CPU @ 3.8GHz and 64.0GB RAM @ 3200MHz. Measured timing values are dedicated to each episode and then summed up over all episodes and finally over all iterations. Compared to the methods such as team Q-learning which require one single learner with a huge Q-table and state-action values, this method that employs multi learners with no correlation has much lower time complexity. We should note that the time complexity of the RL-based algorithm is a function of the size of the gird within each operation region. Therefore, the optimal size of the regions and the number of grids inside each region are determined in a such a way to result in a manageable time to fill out the Q-table for each region. The size of the Q-table is defined based on the number of cells in each region and the number of

TABLE 2. Simulation parameters and required times for the convergence with 5 UAVs and 5 movement actions.

Grid Size:	9×9 (81)	16×16 (256)	27×27(729)	32×32 (1024)		64×64 (4096)			81×81 (6561)	
# of Regions:	9	16	9	64	16	256	64	16	81	9
Region Size:	3×3	4×4	9×9	4×4	8×8	4×4	8×8	16×16	9×9	27×27
# of States:	9	16	81	16	64	16	64	256	81	729
Q-Table Size:	45	80	405	80	320	80	320	1280	405	3645
# of Iterations:	20	20	20	20	20	20	20	20	20	20
# of Episodes:	40	40	40	40	40	40	40	40	40	40
# of Steps:	75	125	600	125	500	125	500	2000	600	6000
Epoch Time(S):	0.16955	0.27642	1.30856	0.25871	1.08395	0.25251	1.09411	4.28103	1.28783	13.04665
Iteration Time(S):	6.78	11.05	52.34	10.34	43.35	10.10	43.76	171.24	51.51	521.86
Total time:	2.26m	3.68m	17.4m	3.44m	14.4m	3.36m	14.58m	57.08m	17.17m	2.89h



(a) 3D topology before regions assignment

(b) 2D topology before regions assignment

(c) 2D topology after regions assignment

FIGURE 6. Topology in 2D and 3D for 5 UAVs in a 64 × 64 grid size and 16 regions.

actions as (*region_size*) × (Number of actions). The Q-tables for different regions are filled out in a parallel manner by the assigned UAVs to these regions. Therefore, increasing the number of regions does not increase the required time to execute the overall mission. The required time for the algorithm under several scenarios based on the grid size and the number of regions is presented in Table 2. For instance, let us consider the region size of 4 × 4 for two cases of grid sizes of 16 × 16 and 64 × 64. While in these two cases, the size of the grids is different, the number of states and the number of actions are equal. As a result, the total simulation times for both cases are almost the same. Therefore, we can keep the grid size in a range to offer an acceptable time complexity by selecting an optimal dimension for the grids as well as the number of the regions.

The initial random locations for 5 UAVs and fixed locations for the primary users and the emergency center in a constant altitude are shown in Figure 6. Figure 6a shows the 3 dimensional topology for all UAVs with a random initialization, and Figure 6b demonstrates the initial point but in a Z plane. Finally, Figure 6c shows the 2D topology for the UAV after the emergency center made decision about the region assignments. Red, blue, and green nodes specify the emergency, primary, and UAV users, respectively.

A. PART-1

In the first part of the simulation, we evaluated the performance of the proposed method with 5 UAVs in a 81 × 81 grid plane with 9 predefined regions. Each region has 27 cells in length and 27 cells in width which means it covers an area of 729 cells that is equal to the number of states for each UAV.

The sum utility for all UAVs is shown in Figure 7. In this sample scenario, UAV [4] is the relay UAV and the rest of them are the sensing UAVs. The summation throughput for each episode is measured which is the accumulated rate for 6000 transmissions during each episode. Based on the plot observation, the UAVs start learning the optimal locations for transmission as the steps proceed. At the final episodes,

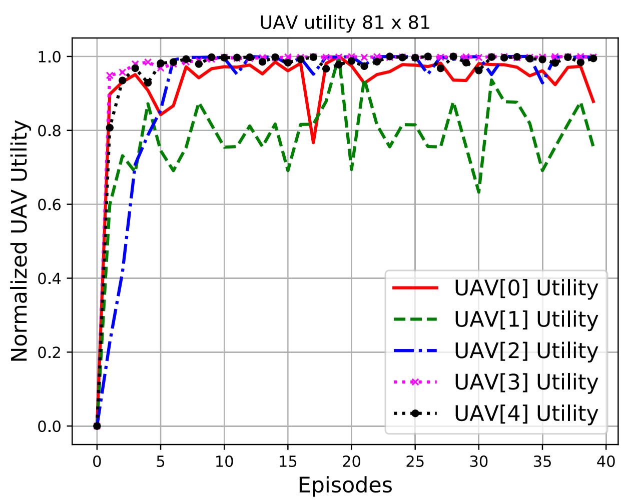


FIGURE 7. Summation throughput rate for all UAVs.

there are more experienced state-action values in the Q-table for each UAV, as a result, they converge faster to the best or optimal location for maximizing the individual throughput. Since the values of throughput depend also on the location of the allocated regions, the plot shows the normalized rate for UAVs to have better demonstration over the behavior.

Figure 8 demonstrates the accumulative reward during each episode for 5 UAVs. As it is shown, the agents start the learning and the exploring at lower episodes and because of that, they gain more random rewards which can be negative or positive. At the final episodes, they are using the updated Q-values based on the reward and Q-value updating functions in (9) and (7), respectively. Hence, it is more beneficial for the UAVs to merge to the optimal locations based on the energy consumption and the rate differential values between two consecutive steps. Hence, they will earn more positive reward compared to the initial episodes. Also, there is no correlation between the UAVs and they follow the same reward function for updating, as a result, they show the same behavior for the accumulative reward.

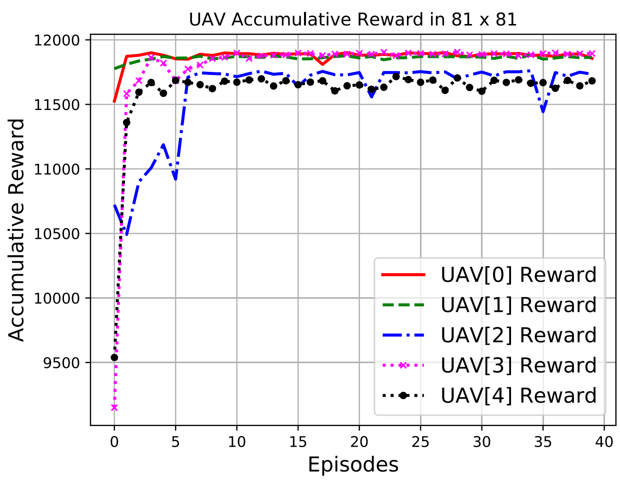


FIGURE 8. **The accumulative reward for all UAVs over all episodes.**

We explicitly defined the energy consideration in the reward function. Hence, the expectation is to have less mobility at the final episodes. Since the mobility costs lots of energy, it brings negative reward for each UAV. The agents learn during action-decisions to have only necessary movements and those necessary moves will result in an optimal location for higher throughput. Based on Figure 9, at the beginning of the learning process, the Q-values are initialized with zeros. Hence, there is no real difference between the exploring and the exploiting and the agents have more mobility to get experience with different cells in their region. At the final episodes, when the Q-table is updated with recent values, the UAVs know their best actions to take in order to reach the optimal states. As a result, they are more stable for the final episodes. At the early episodes, most UAVs took around 2000 actions for mobility, while at the final episodes this number is less than 500 which shows their intention to be more stable.

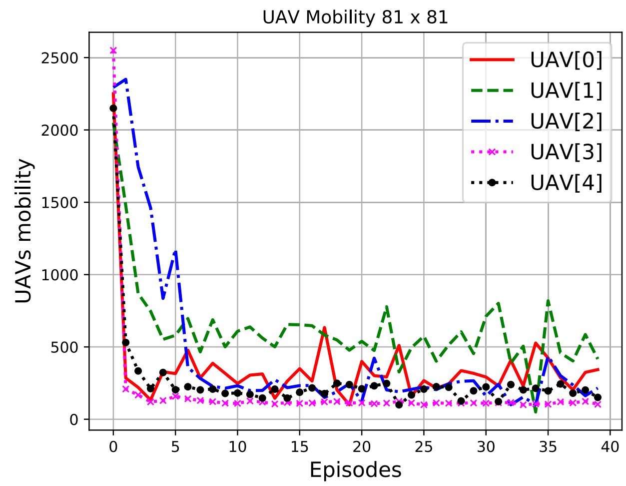


FIGURE 9. **Number of movements for all UAVs in each episode.**

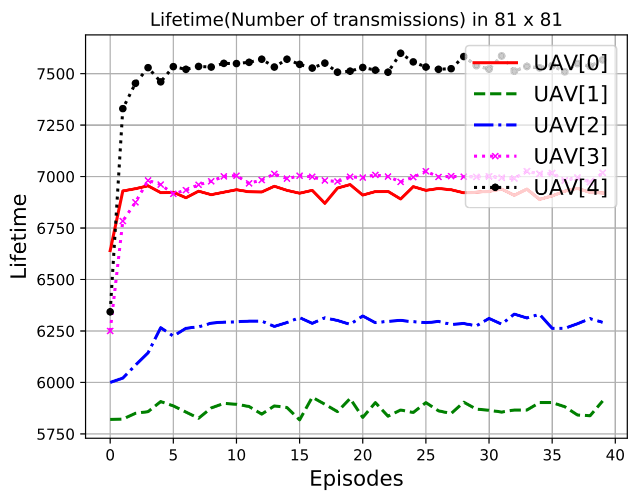


FIGURE 10. **Number of successful transmissions for each UAV over all episodes.**

Figure 10 investigates the lifetime for all UAVs based on the number of successful transmissions. If the battery energy is drained completely, then the UAV is considered as a dead node which cannot operate anymore. In this scenario, instead of 6000 steps, the UAVs transmit or relay infinitely to find the maximum number of transmission. Based on Figure 10, at the early episodes, because of more movements, the UAVs are more prone to higher energy consumption rates and at the final episodes, they are more stable in their locations and the only cost for energy is the transmission. Besides, since the sensing network is borrowing the spectrum from the primary network, the goal is to allocate a UAV which has the longest lifetime to the primary network to utilize the leased spectrum for the critical situation. Based on the emergency center search algorithm, UAV [4] was the most suitable one to relay and it has the best lifetime among all UAVs in the network.

Figure 11 shows the energy consumption rate for all UAVs. To measure the energy consumption rate, we considered the initial energy, the energy at 75% of all steps in each

TABLE 3. Comparison of the required times for convergence between the proposed method and the team-learning approach developed in [27] for a network of 2 UAVs with 6 actions.

Grid Size	2×2 (4)	3×3 (9)	4×4(16)	5×5 (25)	6×6 (36)	8×8 (64)	10×10 (100)
Number of States	4×4=16	9×9=81	16×16=256	625	1,296	4,096	10,000
Size of Q-Table	576	2,916	9,216	22,500	46,656	147,456	360,000
Number of Iterations	10	10	10	10	10	10	10
Number of Episodes in Each Iteration	200	200	200	200	200	200	200
Number of Steps in Each Episode	120	800	2,000	5,000	12,000	20,000	30,000
Process Time for Each Epoch (Sec):	0.102628	0.762713	2.425558	8.757846	49.007077	146.936207	291.176670
Total Required Time:	205.256s	25.4m	80.85m	4.8h	27.2h	3 days	6 days

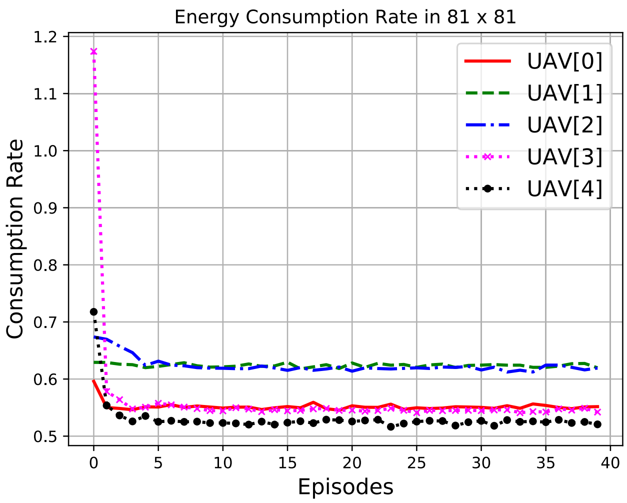


FIGURE 11. Energy consumption rate for all UAVs in each episode.

episode, and the lifetime of the agent in that episode. Based on the figure, UAV [4] has the least energy consumption rate among all UAVs since it performs the task as a relay for the primary (terrestrial) network. The purpose of the searching algorithm for the primary region was to allocate the longest lifetime UAV to the primary network.

Unlike the exponential behavior in [27], the processing time in this work is not increasing at the same pace because of two reasons: first) there is no correlation between the UAVs' strategies and each UAV acts independently, actions and rewards are unique for each UAV, the behavior of one agent does not affect the other ones. Second) increasing the whole grid size does not necessarily increase the processing time since the processing time depends on the number of regions and the regions size. For instance, based on Table. 2, the region size of 4×4 in both grid sizes of 32×32 and 64×64 has the same total simulation time to fill up the Q-table. Using the emergency center to allocate the task among the UAVs reduces the converging time significantly. Table 3 shows the convergence time for the team Q-learning algorithm proposed in [27]. This method works in a fully distributed manner. As a result, training the Q-table requires more time compared to our proposed hybrid method. For instance, it took 6 days to fill the Q-table for 2 UAVs and 6 actions in a 10×10 grid. In our model, it takes almost 3 hours to fill all Q values for 5 UAVs with 5 movement actions in a 81×81 grid plane.

B. PART-2

In the second part of the simulation, we compared the performance of the proposed algorithm with four other methods. We considered 5 modes; mode[0] which is the proposed method of this work that the emergency center considers the priority for critical regions and allocates the UAVs to the regions to prolong the lifetime of the network. The UAVs also use the Q-learning algorithm to find the best cell in their own region to have an optimal throughput rate. In mode[1], the emergency center still considers prioritized regions for the UAVs to have optimal throughput. However, the UAV allocation is random, the UAVs still utilize the RL to find the best cell in the regions. In mode[2], both the regions assignment and the UAV allocations are random; but the UAVs use the RL algorithm in each region. Mode[3] is the opposite of mode[0], the regions are chosen randomly not based on the priority, but the emergency center allocates the UAVs wisely based on the energy consumption rate to maximize the network lifetime. The emergency center considers the prioritized regions and allocates the UAVs to the regions in mode[4] based on the two mentioned search algorithms in Section IV-A. However, the mobility pattern for the UAVs is based on a predefined random path.

Figure 12 demonstrates the summation utilities for 5 different modes. In cases 0 and 1, the optimal summation throughput is derived when the emergency center considers the critical region as a priority for the UAVs. The emergency center also allocates the best UAVs for the sake of network lifetime in case 0. However, the UAVs are chosen randomly in case 1 which does not affect the throughput. On the other hand, in cases 2 and 3, the throughput is decreased since the regions are chosen randomly still the emergency center allocated the UAVs wisely based on the bipartite graph matching. Obviously, case 4 with all random actions has the worst performance.

Figure 13 shows the lifetime for all cases. Case 0, the proposed method, has the longest lifetime in all episodes. To define the lifetime, we consider the relay UAV lifetime, and we assumed that after the battery depletion, the sensing UAVs do not have access to the spectrum anymore. Based on this observation, cases 0 and 3 where the emergency center chooses the optimal UAVs, the lifetime is significantly enhanced compared to other cases. Moreover, choosing the regions also affects the lifetime. It shows that the emergency

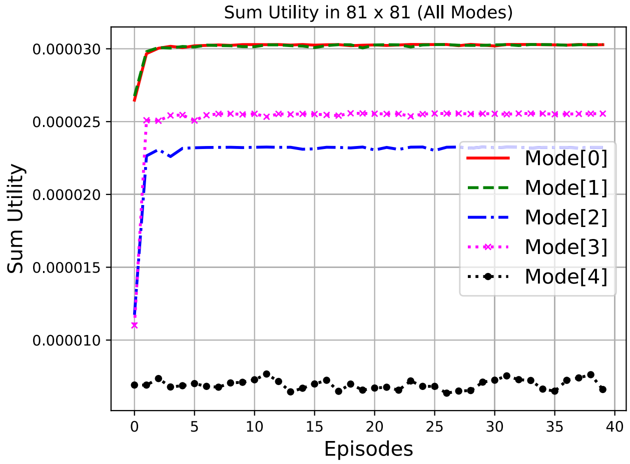


FIGURE 12. Throughput rate comparison in different conditions, Modes[0-4].

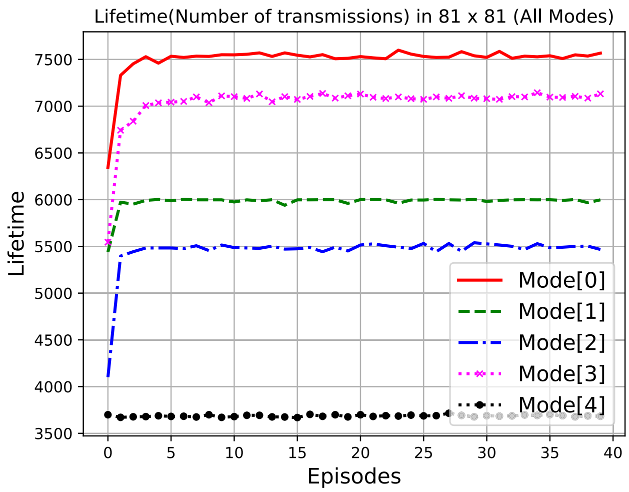


FIGURE 13. Lifetime comparison in different conditions, Modes[0-4].

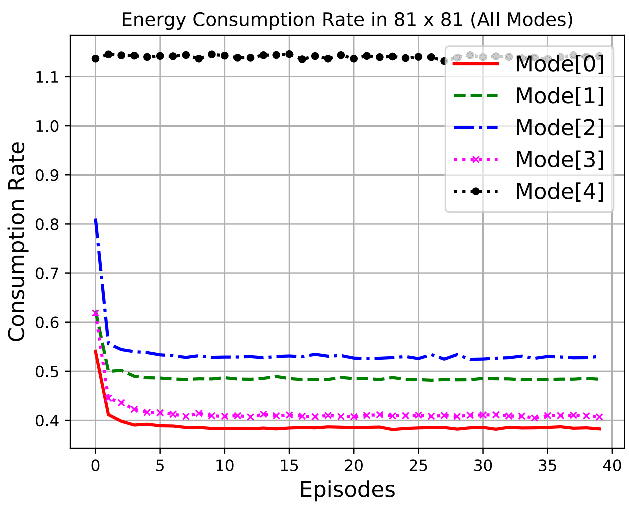


FIGURE 14. Comparison of energy consumption rate in different conditions, Modes[0-4].

center's decisions have crucial results on the lifetime rather than the throughput.

For the energy consumption rate, we considered each UAV's initial energy, the energy at the 75% steps in each episode, and the lifetime of the relay UAV. The average is performed over all the UAVs and the runs. Figure 14 shows the proposed method (case 0) has the least significant energy consumption rates for all the UAVs in the network. Case 3 with the random regions and the allocated UAVs follows case 0. Case 1 with the random UAVs and the regions selected by the emergency center is in the third rank and case 2 with the random regions, the random UAV allocation, and the RL algorithm is the fourth one. Case 4 where both the UAVs and the regions are allocated by the emergency center but with a predefined trajectory has the most energy consumption rate.

VI. CONCLUSION

In this paper, a disaster relief situation is considered as a sample scenario where a group of UAVs observes critical information such as wildfire situations for an emergency center. The UAV network during such critical missions may require additional spare spectrum. To address this demand, we developed a spectrum sharing model where one UAV acts as a relay and forwards data for a terrestrial network in exchange for the required spectrum and the rest of the UAVs utilize this spectrum to transmit the sensed information. The emergency center determines the regions for the primary and secondary networks to maximize the throughput and it allocates the UAVs into the chosen regions based on a predefined grid plane to maximize the network lifetime and reduce the energy consumption. A reinforcement learning approach is used for all the UAVs to find the best cell in each region without any prior information of the environment. The simulation results show that after a certain amount of iterations and episodes, the UAVs converge to the optimal state with a fewer number of actions. We also compared the proposed method with other random assignments and allocations. For larger gird size planes, it is possible to consider the approximate location of the grid plane for hazardous and dangerous areas and perform offline learning to save the time and come up with a predefined Q-table. Then, the UAVs can act completely greedy based on the Q-table to travel to the best location without any extra action.

ACKNOWLEDGMENT

The authors acknowledge the U.S. Government's support in the publication of this paper. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US government or AFRL. Distribution A: Approved for Public Release, distribution unlimited. Case Number 88ABW-2019-5665 on 20 Nov. 2019.

REFERENCES

[1] (Feb. 2017). *Gartner Says Almost 3 Million Personal and Commercial Drones Will be Shipped in 2017*. [Online]. Available: http://www.gartner.com/newsroom/id/3602317

[2] S. Mousavi, F. Afghah, J. D. Ashdown, and K. Turck, "Use of a quantum genetic algorithm for coalition formation in large-scale UAV networks," *Ad Hoc Netw.*, vol. 87, pp. 26–36, May 2019.

[3] G. Sklivanitis, A. Gannon, K. Tountas, D. A. Pados, S. N. Batalama, S. Reichhart, M. Medley, N. Thawdar, U. Lee, J. D. Matyjas, S. Pudlewski, A. Drozd, A. Amanna, F. Latus, Z. Goldsmith, and D. Diaz, "Airborne cognitive networking: Design, development, and deployment," *IEEE Access*, vol. 6, pp. 47217–47239, 2018.

[4] S. Piao, Z. Ba, L. Su, D. Koutsonikolas, S. Li, and K. Ren, "Automating CSI measurement with UAVs: From problem formulation to energy-optimal solution," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2019, pp. 2404–2412.

[5] J. Sánchez-García, J. M. García-Campos, M. Arzamendia, D. G. Reina, S. L. Toral, and D. Gregor, "A survey on unmanned aerial and aquatic vehicle multi-hop networks: Wireless communications, evaluation tools and applications," *Comput. Commun.*, vol. 119, pp. 43–65, Apr. 2018.

[6] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2624–2661, 4th Quart., 2016.

[7] E. Yanmaz, S. Yahyanejad, B. Rinner, H. Hellwagner, and C. Bettstetter, "Drone networks: Communications, coordination, and sensing," *Ad Hoc Netw.*, vol. 68, pp. 1–15, Jan. 2018.

[8] I. Stanojev, O. Simeone, Y. Bar-Ness, and T. Yu, "Spectrum leasing via distributed cooperation in cognitive radio," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 3427–3431.

[9] F. Afghah, M. Costa, A. Razi, A. Abedi, and A. Ephremides, "A reputation-based Stackelberg game approach for spectrum sharing with cognitive cooperation," in *Proc. 52nd IEEE Conf. Decis. Control*, Dec. 2013, pp. 3287–3292.

[10] N. Namvar and F. Afghah, "Spectrum sharing in cooperative cognitive radio networks: A matching game framework," in *Proc. 49th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2015, pp. 1–5.

[11] J. Li, H. Zhao, H. Wang, F. Gu, J. Wei, H. Yin, and B. Ren, "Joint optimization on trajectory, altitude, velocity, and link scheduling for minimum mission time in UAV-aided data collection," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1464–1475, Feb. 2019.

[12] H. Zhao, H. Wang, W. Wu, and J. Wei, "Deployment algorithms for UAV airborne networks toward on-demand coverage," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2015–2031, Sep. 2018.

[13] H. Wang, H. Zhao, W. Wu, J. Xiong, D. Ma, and J. Wei, "Deployment algorithms of flying base stations: 5G and beyond with UAVs," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10009–10027, Dec. 2019.

[14] J. Li, H. Zhao, A. S. Hafid, J. Wei, H. Yin, and B. Ren, "A bio-inspired solution to cluster-based distributed spectrum allocation in high-density cognitive Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9294–9307, Dec. 2019.

[15] D. G. Reina, T. Camp, A. Munjal, and S. L. Toral, "Evolutionary deployment and local search-based movements of 0th responders in disaster scenarios," *Future Gener. Comput. Syst.*, vol. 88, pp. 61–78, Nov. 2018.

[16] D. G. Reina, H. Tawfik, and S. L. Toral, "Multi-subpopulation evolutionary algorithms for coverage deployment of UAV-networks," *Ad Hoc Netw.*, vol. 68, pp. 16–32, Jan. 2018.

[17] C. Zhang and W. Zhang, "Spectrum sharing for drone networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 1, pp. 136–144, Jan. 2017.

[18] M. McHenry, Y. Zhao, and O. Haddadin, "Dynamic spectrum access radio performance for UAS ISR missions," in *Proc. Mil. Commun. Conf. (MILCOM)*, Oct. 2010, pp. 2345–2350.

[19] J. Kakar and V. Marojevic, "Waveform and spectrum management for unmanned aerial systems beyond 2025," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–5.

[20] H. Wang, J. Wang, G. Ding, J. Chen, Y. Li, and Z. Han, "Spectrum sharing planning for full-duplex UAV relaying systems with underlaid D2D communications," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1986–1999, Sep. 2018.

[21] J. Lyu, Y. Zeng, and R. Zhang, "Spectrum sharing and cyclical multiple access in UAV-aided cellular offloading," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[22] T. X. Brown and S. Jaroonvanichkul, "Policy-based radios for UAS operations," in *Proc. IEEE Globecom Workshops*, Dec. 2012, pp. 1614–1619.

[23] T. Brown and D. Sicker, "Secure distribution of spectrum policies to cognitive radios in unmanned aircraft systems," in *Proc. AIAA Infotech@Aerospace*, Apr. 2010, p. 3333.

[24] S. Mousavi, F. Afghah, J. D. Ashdown, and K. Turck, "Leader-follower based coalition formation in large-scale UAV networks, a quantum evolutionary approach," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 882–887.

[25] C. Guan, A. Mohaisen, Z. Sun, L. Su, K. Ren, and Y. Yang, "When smart TV meets CRN: Privacy-preserving fine-grained spectrum access," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2017, pp. 1105–1115.

[26] A. Shamsoshoara, M. Khaledi, F. Afghah, A. Razi, and J. Ashdown, "Distributed cooperative spectrum sharing in UAV networks using multi-agent reinforcement learning," in *Proc. 16th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2019, pp. 1–6.

[27] A. Shamsoshoara, M. Khaledi, F. Afghah, A. Razi, J. Ashdown, and K. Turck, "A solution for dynamic spectrum management in mission-critical UAV networks," 2019, *arXiv:1904.07380*. [Online]. Available: http://arxiv.org/abs/1904.07380

[28] Y. Ben-Asher, S. Feldman, P. Gurfil, and M. Feldman, "Distributed decision and control for cooperative UAVs using ad hoc communication," *IEEE Trans. Control Syst. Technol.*, vol. 16, no. 3, pp. 511–516, May 2008.

[29] J. J. Acevedo, B. C. Arrue, J. M. Diaz-Banez, I. Ventura, I. Maza, and A. Ollero, "Decentralized strategy to ensure information propagation in area monitoring missions with a team of UAVs under limited communications," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, May 2013, pp. 565–574.

[30] M. Carpin, S. Rosati, M. Emtiyaz Khan, and B. Rimoldi, "UAVs using Bayesian optimization to locate WiFi devices," 2015, *arXiv:1510.03592*. [Online]. Available: http://arxiv.org/abs/1510.03592

[31] R. C. Palat, A. Annamalai, and J. H. Reed, "Cooperative relaying for ad-hoc ground networks using swarm UAVs," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2005, pp. 1588–1594.

[32] R. Fan, J. Cui, S. Jin, K. Yang, and J. An, "Optimal node placement and resource allocation for UAV relaying network," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 808–811, Apr. 2018.

[33] Y. Chen, W. Feng, and G. Zheng, "Optimum placement of UAV as relays," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 248–251, Feb. 2017.

[34] S. D. Muruganathan, X. Lin, H.-L. Maattanen, J. Sedin, Z. Zou, W. A. Hapsari, and S. Yasukawa, "An overview of 3GPP release-15 study on enhanced LTE support for connected drones," 2018, *arXiv:1805.00826*. [Online]. Available: http://arxiv.org/abs/1805.00826

[35] Q. Sun, D. C. Cox, H. C. Huang, and A. Lozano, "Estimation of continuous flat fading MIMO channels," in *Proc. IEEE Wireless Commun. Netw. Conf. Record. (WCNC)*, vol. 1, Mar. 2002, pp. 189–193.

[36] S. Coleri, M. Ergen, A. Puri, and A. Bahai, "Channel estimation techniques based on pilot arrangement in OFDM systems," *IEEE Trans. Broadcast.*, vol. 48, no. 3, pp. 223–229, Sep. 2002.

[37] F. Afghah, A. Shamsoshoara, L. Njilla, and C. Kamhoua, "A reputation-based Stackelberg game model to enhance secrecy rate in spectrum leasing to selfish IoT devices," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 312–317.

[38] A. Shamsoshoara and Y. Darmani, "Enhanced multi-route ad hoc on-demand distance vector routing," in *Proc. 23rd Iranian Conf. Elect. Eng.*, May 2015, pp. 578–583.

[39] F. Shah-Mohammadi and A. Kwasinski, "Deep reinforcement learning approach to QoE-driven resource allocation for spectrum underlay in cognitive radio networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[40] F. S. Mohammadi and A. Kwasinski, "QoE-driven integrated heterogeneous traffic resource allocation based on cooperative learning for 5G cognitive radio networks," in *Proc. IEEE 5G World Forum (5GWF)*, Jul. 2018, pp. 244–249.

[41] M. R. Souryal and B. R. Vojcic, "Performance of amplify-and-forward and decode-and-forward relaying in Rayleigh fading with turbo codes," in *Proc. IEEE Int. Conf. Acoust. Speed Signal Process.*, vol. 4, May 2006, p. 4.

[42] G. Levin and S. Loyka, "Amplify-and-forward versus decode-and-forward relaying: Which is better?" in *Proc. 22th Int. Zürich Seminar Commun. (IZS)*, 2012, pp. 1–5.

[43] Z. Guan, T. Melodia, D. Yuan, and D. A. Pados, "Distributed resource management for cognitive ad hoc networks with cooperative relays," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1675–1689, Jun. 2016.

[44] F. Chen, W. Su, D. A. Pados, J. D. Matyjas, and M. J. Medley, "Relay location optimization for differential amplify-and-forward cooperative relaying," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–7.

[45] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.
[46] M. Yu, J. Li, and H. Sadjadpour, "Amplify-forward and decode-forward: The impact of location and capacity contour," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2005, pp. 1609–1615.
[47] A. Nosratinia, T. E. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 74–80, Oct. 2004.
[48] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 237–285, Jan. 1996.
[49] S. S. Mousavi, M. Schukat, E. Howley, and P. Mannion, "Applying q (λ))-learning in deep reinforcement learning to play Atari games," in *Proc. AAMAS Adapt. Learn. Agents (ALA) Workshop*, 2017, pp. 63–68.
[50] S. S. Mousavi, M. Schukat, and E. Howley, "Deep reinforcement learning: An overview," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2016, pp. 426–440.
[51] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[52] F. S. Melo, "Convergence of Q-learning: A simple proof," Inst. Syst. Robot., Lisbon, Portugal, Tech. Rep. 4, 2001.
[53] T. Jaakkola, M. I. Jordan, and S. P. Singh, "Convergence of stochastic iterative dynamic programming algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 703–710.
[54] P. Dayan and C. Watkins, "Q-learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, 1992.

ALIREZA SHAMSOSHOARA received the B.Sc. degree in electrical engineering from Shahid Beheshti University (SBU), Tehran, Iran, and the M.Sc. degree in electrical and communication engineering from the Khaje Nasir Toosi University of Technology (KNTU), Tehran, in 2012 and 2014, respectively, and the M.Sc. degree in informatics from Northern Arizona University, in 2018. He is currently pursuing the Ph.D. degree with the School of Informatics, Computing and Cyber Systems, Northern Arizona University. He worked as a Network Engineer, from 2015 to 2017. His main research interests include wireless networks, UAV networks, spectrum sharing, and machine learning.

FATEMEH AFGHAH was an Assistant Professor with the Electrical and Computer Engineering Department, North Carolina A&T State University, Greensboro, NC, USA, from 2013 to 2015. She is an Assistant Professor with the School of Informatics, Computing and Cyber Systems, Northern Arizona University (NAU), Flagstaff, AZ, USA, where she is the Director of Wireless Networking and Information Processing (WiNIP) Laboratory. Her research interests include wireless communication networks, decision making in multiagent systems, radio spectrum management, hardware-based security, and artificial intelligence in healthcare. She received the Air Force Office of Scientific Research Young Investigator Award, in 2019 and the NSF CRII Award, in 2017.

ABOLFAZL RAZI received the B.S. degree from Sharif University, in 1998, the M.S. degree from Tehran Polytechnic, in 2001, and the Ph.D. degree from the University of Maine, in 2013, all in electrical engineering. He has held two postdoctoral positions in the field of machine learning and predictive modeling at Duke University, from 2013 to 2014, and Case Western Reserve University, from 2014 to 2015. He is an Assistant Professor of electrical engineering with Northern Arizona University (NAU). His current research interests include AI-enabled wireless networking, nonlinear compressive sensing, topology control, UAV networks, and predictive modeling. He was a recipient of several competitive awards, including the Best Research of MCI, in 2008, the Best Graduate Research Assistant of the Year Award from the College of Engineering, University of Maine, in 2011, and the Best Paper Award from the IEEE/CANEUS Fly by Wireless Workshop, in 2011.

SAJAD MOUSAVI received the B.Sc. degree in computer engineering from Zanjan University (ZNU), Zanjan, Iran, in 2010, and the M.Sc. degree in artificial intelligence and robotics from the Iran University of Science and Technology (IUST), Tehran, Iran, in 2012. He is currently pursuing the Ph.D. degree with the School of Informatics, Computing and Cyber Systems, Northern Arizona University. He worked as a Research Assistant in machine learning and deep learning with the National University of Ireland Galway, Galway, Ireland, from 2015 to 2017. His main research interests include machine learning, deep learning, computer vision, multiagent systems, and task allocation.

JONATHAN ASHDOWN was born in Niskayuna, NY, USA, in 1984. He received the B.S., M.S., and Ph.D. degrees from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2006, 2008, and 2012, respectively, all in electrical engineering. His Ph.D. dissertation was on a high-rate ultrasonic through-wall communication system using MIMO-OFDM in conjunction with interference mitigation techniques. In 2012, he was a recipient of the Best Unclassified Paper Award at the IEEE Military Communications Conference. From 2012 to 2015, he worked as a Civilian Research Scientist with the Department of Defense (DoD), SPAWAR Systems Center Atlantic, Charleston, SC, USA, where he was involved in several basic and applied research projects for the U.S. Navy, mainly in the area of software defined radio. In 2015, he transferred within DoD. He is currently an Electronics Engineer with the Air Force Research Laboratory, Rome, NY, USA, where he is involved in the research and development of advanced emerging technologies for the U.S. Air Force.

KURT TURK is a Communications Program Manager for the Air Force Research Laboratory, Rome, NY, USA, with over 20 years of experience focused on the development and integration of innovative and affordable information transmission and reception technologies to enhance warfighting utility, survivability, and mission assurance. His areas of expertise are wireless communications, mobile ad-hoc networking, cross layer communications, airborne networking, and modeling and simulation. In the past 16 years, he has worked in networking and communications for the U.S. Government. Some projects include OPNET modeling for directional tactical radios, bio-inspired networking, and spectrum mapping.